OPEN ACCESS
Vietnam Journal of Computer Science
Vol. 8, No. 4 (2021)

DOI: 10.1142/S2196888821500226

World Scientific
www.worldscientific.com

Online Tracking: When Does it Become Stalking?

Bede Ravindra Amarasekara*, Anuradha Mathrani† and Chris Scogings‡

School of Natural and Computational Sciences
Massey University, New Zealand
**b.amarasekara@massey.ac.nz*
†*a.s.mathrani@massey.ac.nz*
‡*c.scogings@massey.ac.nz*

Received 30 June 2020
Accepted 24 November 2020
Published 25 May 2021

Online user activities are tracked for many purposes. In e-commerce, cross-domain tracking is used to quantify and pay for web-traffic generation. Our previous research studies have shown that HTTP cookie-based tracking process, though reliable, can fail due to technical reasons, as well as through fraudulent manipulation by traffic generators. In this research study, we evaluate which of the previously published tracking mechanisms are still functional. We assess the efficacy and utility of those methods to create a robust tracking mechanism for e-commerce. A failsafe and robust tracking mechanism does not need to translate into further privacy intrusions. Many countries are rushing to introduce new regulations, which can have a negative impact on the development of robust technologies in an inherently stateless eco-system. We used a multi-domain, purpose-built simulation environment to experiment common tracking scenarios, and to describe the parameters that define the minimum tracking requirement use-cases, and practices that result in invading privacy of users. This study will help practitioners in their implementations, and policy developers and regulators to draw up policies that would not curtail the development of robust tracking technologies that are needed in e-commerce activities, while safeguarding the privacy of internet users.

Keywords: Cross-domain; tracking; affiliate marketing; HTTP cookie; XDT.

1. Introduction

While HTTP cookies have been providing reliable tracking capabilities for over two decades,[1] previous research studies have exposed underlying issues where HTTP cookie-based tracking mechanism can fail.[2] There are also instances where fraudulent parties can manipulate tracking systems to falsify tracking data,[3–5] usually for monetary gain. In recent years research findings have presented alternative methods

*Corresponding author.

for state management, specifically those that can be extended as tracking methods.[6–11] Traditional tracking methods such as HTTP cookies, which have been specifically developed to manage state, have been in use for tracking for a long time and are likely to remain usable even in the future. Any future developments can be expected to remain backward compatible or further enhanced, as they are meant for state management purpose. In contrary, the newer alternative methods presented in recent research studies may usually have a shorter lifespan and cannot be guaranteed to be usable over time. As those technologies evolve to serve their intended purposes in future, they can lose their usability as a tracking method. In this research, we examined some of the newer tracking methods presented in previous studies and tested their current usability and whether they can complement existing technologies to improve the robustness of the tracking process within an e-commerce environment. That would mean an improved and fail-safe cross-domain tracking capability for e-commerce.

Nevertheless, an improved robustness and accuracy, may appear to be a more persistent and privacy invasive threat, in the minds of some privacy advocates. Online tracking is fast becoming synonymous with stalking, with increasing number of countries rushing to introduce plethora of new privacy laws. Adhering to multitude of regional and country specific privacy laws on the Internet where physical borders are obscure, and compliance with such regulations is not only difficult, but also is somewhat defeating the purpose of such privacy concerns.[12] New research findings suggest General Data Protection Regulation.[13] (GDPR) introduced by European Union as recently as May 2018, does not achieve its intended purpose, due to click-fatigue.[14] While it is important to protect the privacy of internet users, it is equally important to develop and maintain robust mechanisms to maintain state in a traditionally stateless ecosystem, across geographically distributed multiple domains, making e-commerce activities reliable. Therefore, it necessitates identifying and categorizing different use-cases of cross-domain user tracking on the internet. Such tracking practices span from a purely technological necessity in one end to person-identifying and data-marketing endeavors at the opposite extremity. This segmentation enables practitioners and regulators to define and adhere to regulations and best practices, that would effectively curb privacy intrusions without unintended consequences of technological curtailments. This paper examines different technologies that may be used to strengthen the online tracking process, thereby also verifying which of the previously presented technologies are still usable for tracking purpose today, with current developments in technology. Then, it examines different use-cases of online tracking and categorizes them into levels of privacy intrusion involved and levels of indispensability in terms of a technical necessity. Finally, this paper presents how improved and more reliable online tracking techniques can enhance e-commerce activity without compromising privacy of internet users when used purely as an underlying technology. This paper also reveals which techniques have what levels of intrusions, when combined with Person Identifying Information (PII). This knowledge will provide clarity to policy developers and legislature to

formulate effective and consistent regulations and policies without undermining the technical necessities of legitimate e-commerce activities. It will also facilitate practitioners to define boundaries in their implementations. Importantly, the scientific community can extend this research to develop technological solutions and frameworks that can automate machine-to-machine negotiation processes, protocols and standards between client and server while adhering to privacy guidelines, thus eliminating human intervention that leads to "click-fatigue".[14]

The topic related to "Improving the robustness of the tracking process" was discussed in our previous paper presented at ACIIDS 2020 conference.[15] This paper extends our discussion further with privacy concerns that are associated with online tracking, in the given context.

2. Related Literature

Hypertext Transport Protocol (HTTP) is stateless by design. Application "state" is not maintained between calls to an HTTP server and every call is considered a new request. With the development of the Internet and e-commerce activities, a mechanism to manage state was required, and HTTP-cookie was introduced.[1] Using hidden fields on the page and embedding parameters within the request URL are some of the other state management methods used. Most e-commerce applications need a persistent state management mechanism, as it is vital to "remember" choices that individual customers make, and information that they enter into web forms, as they navigate through webpages on a site, before they submit the form to complete a transaction. Saving the customer's choice of language, currency type and other frequently used choices beyond that single transaction and using them to pre-fill a form enhances customer satisfaction. While customers gain a positive user experience, businesses gain the ability to transform behavioral data that can reflect customer habits and preferences, which are then used for targeted marketing and business analytics. We present scenarios which will enable practitioners and regulators to define boundaries between user experience, technical necessity and privacy intrusion.

With the introduction of "Local Storage"[11] with HTML5, another reliable mechanism that can store data locally within a client browser has been made available to web applications.[6,10,11] As a client-side technology, the web server cannot interact with the "Local Storage" directly; all interactions are managed by JavaScript. Usually, a unique identifier for each visitor is stored in the "Local Storage", that allows the web server to recall the customer related information stored in the web server, using this unique identifier. "Local Storage" can be used as a tracking mechanism, though it is less versatile than a HTTP-cookie, not being intended for the tracking purpose.[9,16] With the introduction of "ETag" as a web cache validation mechanism,[17] it was discovered that ETags too can be used as a tracking mechanism.[6]

Existing literature shows that Flash-cookie or the local storage of an adobe flash application, officially named "Local shared objects" has also been successfully used in the past as a "super-cookie"; it is considered to be almost indestructible as it is not

managed by the browser and has been used to re-spawn deleted HTTP cookies.[6,9,10] As per literature, blocking of HTTP-cookies on the browser, deleting of cookies, browser cache or browsing history did not have any effect on the Flash-cookie. Even switching to "in-private" browsing could not disable it either, as it is not part of the browser infrastructure. Its purpose was to provide "Local Storage" to Adobe Flash applications.

2.1. *Cross-domain tracking*

Cross-domain tracking (XDT) involves tracking user-interactions across multiple web domains that may be geographically distributed and owned by different entities that do not communicate directly with each other. XDT capabilities are useful for different purposes. Generating network traffic today, happens across multiple websites. A user may click on a product that appear on one website, that causes the visitor to arrive at the e-commerce site that sells the product. In between, the traffic moves through an intermediary site that records and keeps track of the source and destination of the traffic, as the e-commerce site must pay the source for traffic generation. There can be many intermediaries involved in one e-commerce transaction, where each intermediary needs to be rewarded.[5,18–21] Hence this kind of tracking is a technical necessity, as an underlying technology used in different e-commerce activities.[22] Such tracking capability is achieved using "Cookies" or similar methods, that can store a small amount of data to identify a web-user uniquely, which does not capture Personally Identifiable Information (PII), which therefore is usually not considered to be a privacy threat. The unique identifier is usually a long number or a GUID. The same tracking method can also be used to track web-users for multiple other reasons by commercial and governmental entities. They may capture online behavioral data that is combined with PII to create comprehensive user profiles that invade the privacy of users, without their explicit permission. As both PII and non-PII-based tracking use similar technologies to capture data, regulations that restrict usage of such techniques (e.g. using HTTP cookies) can adversely affect scenarios that use tracking only as an underlying technology to manage state.

Some online tracking scenarios are as follows:

- Affiliate marketing model, which is one of the most cost-efficient online marketing methods available to e-marketing practitioners. It needs the capability to track visitors who are viewing and clicking on advertisements placed on affiliates' websites.[23–26] The tracking mechanism traces clicks and successful outcomes; and pays commissions to affiliates.
- Another usage is for customization web content and personalization of advertisement based on a visitor's historical browsing data.[27] Without this capability, internet users can feel hassled, when products and services that do not even vaguely interest them, appear at most of the websites they visit.[28] Also, the

advertisers will be wasting their marketing budget on audiences that do not yield them any positive outcomes.

- Customer behavioral data within an e-commerce site (e.g. duration spent on site and on specific pages, products perused, success rate, etc.) are useful for a marketer, and can be easily generated within the e-commerce application. By subscribing to an external business analytics provider, such data can be combined with customer demographics obtained through insights over interactions beyond the boundaries of the practitioner, to generate richer person-profiles useful for a marketeer.[18]
- Security establishments use tracking technology to identify people who are deemed a security threat. They are flagged across multitude of websites and their activities are monitored.
- Third party companies such as Cambridge Analytica profiles people with the help of people's social media affiliations and interests. By using such profiling methods, they are capable of undertaking nefarious activities such as influencing and creating biased opinions to manipulate political and election outcomes in many countries around the globe.[29]

2.2. *Privacy concerns related to online tracking*

Most web traffic generation methods involve a minimum of three web domains. For example, organic or paid searches (e.g. with Google) would involve the Google domain, an e-commerce domain, and the visitor domain. Apart from online traffic generation endeavors, business analytics and customer demographic data services also require XDT capability.[30] Usually e-marketing services gather behavioral data on customers, such as origin of the traffic, total vs. successful visit counts, products perused by customer, time duration spent on different pages and other customer demographic information that helps marketers to target marketing campaigns to specific audiences. They also provide helpful insights for a marketeer to understand if the customer needs are met by their product offerings.

If the tracking process is carried out by the e-commerce practitioner in-house, then the available visitor information is limited to the interactions within practitioner's own domain. But as third-party tracking service providers offer services to many e-commerce sites, they can offer additional information for a premium price. Such information could include, e.g. which website did the visitor arrive from, which website did the visitor go to or what products were perused in previous sites, among other useful information. Some service providers offer remarketing leads by using the information they have gathered in competitor sites that have subscribed to the same tracking service. Using a tracking service provider expands the accessibility scope of visitor data but is still limited to those e-commerce sites that have subscribed to the same tracking service provider.

The hierarchical nature of the information access capability of various service providers enables information exploitation to occur at different degrees. As one

traverses up the hierarchical tree, service providers sitting at a higher level have increasingly wider visibility. Services at the top of the hierarchy have visibility over the largest number of node sites. Almost every internet user utilizes some form of a service provided by at least one of the largest global service providers such as Google, Facebook, Microsoft, Apple or similar tech giants. Often a person may be using services of all or most of the above tech giants. Being on top of the hierarchical tree, they have visibility of user-interaction over most of the internet.[31] To use services provided by these tech giants, one needs to create a user profile with personally identifiable information and sign-in with a user account. A cookie that is placed into the site-visitor's web browser during the sign-in process will identify the visitor uniquely across all services offered by these tech giants and at numerous other seemingly independent websites. Often the presence of these tech giants is not directly visible to the visitors of a third-party website. But, unbeknown to the visitor, most third-party websites utilize some services of these tech giants in the background, such as resources from a Content Delivery Network (CDN), widgets or subscription to a business analytics service. When such a resource is loaded to the browser while rendering the third-party web page, the cookie set by the tech giant is automatically sent back to the web server with each new request. That reveals the presence of the user at the specific third-party site, thus allowing such services to gather data on user's navigation across the internet. When using a browser application provided by one of these tech giants, the exposure of the user data increases even further, as the browser can monitor all interactions with websites, without depending on the cookies. Using operating systems or hardware (e.g. phones, tablets) provided by these tech-giant has the highest exposure, as the personally identifiable information are available at the operating system level.[32] Previous research found that 80% of Alexa's top one million websites were being tracked by Google, while another found the percentage to be even higher at 97% among the top 100 websites.[6,27]

Business Analytic services such as Google Analytics (Universal Analytics) offer standard services free of cost to everybody, while charging a price for premium services. The comprehensiveness of the insights sold as premium services depends on their ability to track users across the entire internet.[30,33] Therefore, many such service providers offer free services with limited features to users who are not willing to pay for those services. This in turn will allow a provider to harvest comprehensive set of user related data of a large customer base, that makes up the product which will be marketed as a premium service.

Some of the free services that are offered by such operators are: web browsers, e-mail services, cloud storage, business analytics, widgets such as counters, exchange rate and weather information, CDN services, DNS services and others. The information exploitation mantra is simple: place as many cookies on the client browsers as possible by offering shared resources through CDNs or provide as many free services as possible, since it will enable the service provider to place a cookie, and gather as many "pings" along the way.

2.3. *This study*

This research carried out experiments on an Affiliate Marketing Network (AMN) within an e-commerce scenario. While some large e-commerce practitioners such as e-bay, amazon.com, etc. manage the tracking process in-house, others choose to entrust it to specialist tracking service providers, such as AMNs.

An AMN is a typical example of a large network of affiliates who generate web traffic for e-commerce sites. Affiliates are popular websites based on diverse themes, who already have a large audience of web traffic. They agree to display advertisements of e-commerce sites for a fee. Some advertisers pay affiliates a fee to simply display an advertisement, while others may expect more visitor interactions such as requiring a visitor to click on an advertisement and arrive at the e-commerce site. Yet others pay a commission to affiliates, only if a visitor makes a purchase. A tracking pixel of AMN will be placed on affiliate's webpage, which is usually a small piece of JavaScript, which will cause the visit to be registered on the AMN's tracking server. In case of commission-based advertising, another "conversion-pixel" is placed on the e-commerce site's payment confirmation page, which will cause the AMN's tracking server to register the total prices and the commission amounts due to the affiliate. In spite of the transaction originating and ending at vastly different web domains, possibly over different geographical locations and over a longer time span, the HTTP cookie-based tracking process enables the AMN to accurately recognize the affiliate who displayed the advertisement to the customer and reward the affiliate with the correct amount of commission or fee.[5,18–21]

But there are instances that the cookie-based tracking process can fail.[34] We discuss some of those scenarios and investigate if the HTTP-based tracking process can be made more robust by supplementing the HTTP cookie-based technology with other technologies that we encountered in our previous research work.

3. Methodology

Information systems research falls broadly in two research paradigms: behavioral research and design science research. The purpose of design science research is to solve an existing industry problem by producing design artifacts as outputs.[35–37] Our research aims to solve an existing industry problem on how to make the online cross domain tracking processes more robust while maintaining the tracking process within bounds of technical necessities, thus avoiding privacy intrusions of internet users.

3.1. *Setting up of test environment*

An experiment on cross-domain tracking (XDT) requires multiple domain-based networks on separate IP segments that are interconnected with same network technologies and topologies to simulate internet infrastructure. To track visitor-interactions across multiple domains, all the domains being tracked require the

ability to communicate with a mutually available central tracking domain. From the XDT scenarios discussed above, a simulation of an Affiliate Marketing Networks (AMN) was chosen for this experiment, which comprises a minimum of four separate domains. Such network allows us to test different XDT-based technology implementations. The setup can simulate different e-marketing models such as display advertising, pay-per-click model (PPC) or revenue-sharing models such as cost-per-acquisition (CPA). It can also be used to simulate business analytic services, CDN's and other multi-domain transactions. Bespoke web applications abstracted to the minimum requirements for each category of the four domains were created as part of this research. Virtual servers were used to create a multi-domain network environment for all our experiments as shown in Fig. 1. Each domain-based virtual network consisted of a Primary Domain Controller (PDC), Domain Name Server (DNS), a Web Server, a Database Server. Each domain was connected via virtual network infrastructure that allowed inter-domain routing using TCP-IP protocol. On completion of our experiments we created publicly accessible real-world web domains with the same names, facilitating researchers to executes some of the tests.

The four categories of domains used in this experiment are described later in the chapter. An XDT process starts with an internet user (domain 1: "Customer domain"). When the internet user visits his favorite blog or special interest site (domain 2: "Affiliate"), which also displays third-party banner advertisements of

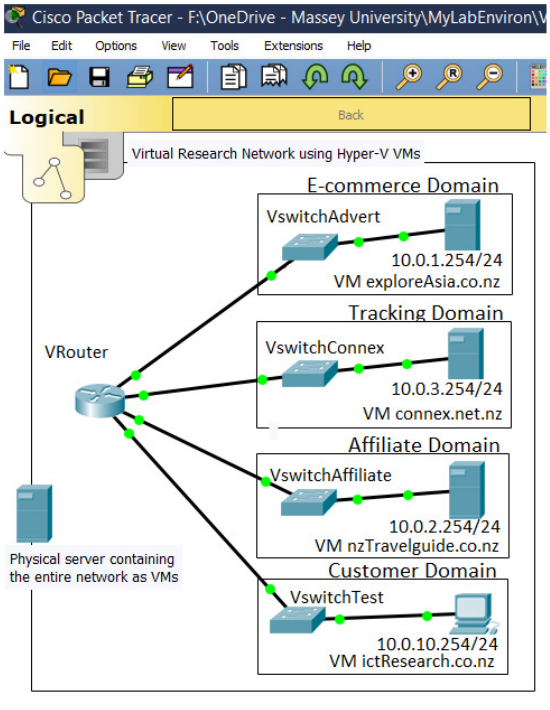

Fig. 1. Virtual Network Diagram.

different products, the visitor clicks on an advertisement. This click is first recorded at a tracking service provider (domain 3: "Tracking domain"). Then it takes the visitor to the e-commerce site that sells the product (domain 4: "e-commerce domain"). In a real-world scenario, one e-commerce site uses more than one affiliate, often in hundreds. Each affiliate has more than one visitor. Also, a tracking services provider usually provides tracking services to more than one e-commerce site. To experiment privacy intrusion and how a tracking provider can track a visitor interaction across all the different e-commerce sites that it provides services to, we need the ability to create multiple instances of visitors, affiliates and e-commerce sites. Using virtual machines, we were able to create as many instances from a master copy of each category of domains. Each category was pre-configured with bespoke software, which we developed as part of this research, to carry out a specific role.

Participating network domains were classified based on their functionalities within an XDT scenario into four groups:

3.1.1. *Tracking domain*

Connex.net.nz domain was configured as a tracking domain, which is at the center of all the tracking activities in this study. The tracking server contained a bespoke software that had the function and ability to track user activities within all other e-commerce domains. "Pixel-codes" embedded in the webpages belonging to e-commerce and e-marketing sites cause visitor-browsers to "ping" the tracking server at connex.net.nz. This enabled us to test tracking service capabilities for AMNs based on different affiliate marketing models, e.g. display advertising, click advertising and revenue-share advertising models. Different service endpoints were created to offer different services which are discussed later in this section.

3.1.2. *E-commerce domains*

Bestcars.ecopng.com, exploreasia.co.nz and ecovillagerundu.com domains were configured as e-commerce servers which subscribed to the tracking services provided by connex.net.nz. Each e-commerce server contained a basic product display page, with a shopping cart functionality. A "conversion-pixel" was placed on each payment confirmation page to track online purchases against "clicks" generated by affiliates. VMs allowed creating multiple instances as needed, and multiple subdomains of the above three domains were used to host the newly created e-commerce servers. Such configurations were required for experiments that needed to observe how tracking information from multiple unrelated e-commerce domains can be shared, how it affects the privacy of web-visitors, and how the technical aspects of XDT can be improved.

3.1.3. *Affiliate domains*

NZtravelguide.org.nz, NewZealandTravel.net.nz domains and multiple sub-domains were configured as affiliates for the above e-commerce sites. Each Affiliate site

contained a landing page, with banner advertisements as "click-pixels. Each affiliate instance was hosted on a different network segment during experiments that required multiple affiliates.

3.1.4. *Internet-user domains*

Computers and mobile devices were added to ICTresearch.co.nz domain representing a multitude of visitors using different devices to access the internet. Devices of this "Test" group were placed outside of all other domains. Devices using different operating systems, different browsers, physical mobile devices, and mobile simulations on desktop browsers were used to repeat each test within different combinations of above variables.

Above bespoke software for the e-commerce sites and for the tracking server were developed as part of this research. This allowed the researchers to add and upgrade functionality during the experiments to suit and incorporate any changing needs.

3.2. *Test setup*

The following test parameters were defined to measure the success of cross-domain tracking capability. Using HTTP-cookies, the following capabilities were ascertained as a baseline for the test environment. Following seven tests were conducted:

Test 1: Loading a page or clicking a banner on any of the tracked pages of the e-marketing domains causes a visit to be accurately registered on the tracking server.

Test 2: The ability for payment confirmation pages of e-commerce sites to accurately and reliably transmit the affiliate identifier and total price of items purchased to the tracking server. These two test capabilities encompass the tracking process needed for an affiliate marketing network.

Test 3: Ability to simultaneously maintain visitor identity between two windows of the same browser.

Test 4: Ability to simultaneously maintain visitor identity between two tabs within the same window of a browser.

Test 5: Despite the "private browsing" mode of a browser, the tracking server has ability to identify a user with a previously saved identifier instead of recording them as a new user.

Test 6: Ability to identify a visitor uniquely when using different browsers within the same device. Usually, browsers do not share cookies, therefore will appear as a new visitor for each browser.

Test 7: Ability to continue to identify a visitor even after the browser cookies are deleted.

3.3. *Privacy intrusion simulations*

Only one instance of a tracking server (connex.net.nz) is required to track visitor interactions across all participating domains and during all different tests. Though

one instance per each of the other three types of domains (e-commerce, affiliate and visitor) is enough to experiment cross-domain tracking functionality, we have extended the experiment by adding multiple instances of each of the three categories. Using multiple e-commerce domains allowed us to simulate a real-world Affiliate Marketing Network (AMN), which provides tracking services to multiple e-commerce sites. It further allowed us to simulate business analytics services such as Google Analytics. The risks associated with rogue Content Delivery Networks (CDN) were experimented using the same.

Different configurations were used to test hierarchical nature of services and associated information exposure. Minimum requirements of non-PII data required for successful tracking was compared against PII data gathered in the process of business analytics gathering.

4. Results

The results of the seven tests show, that "super cookie" concept.[6,10] discussed in previous research do not apply anymore at the same degree. Super cookie concept was not one specific technology, but a combination of technologies, when used in tandem would result in an indestructible tracking solution that can be easily respawned when deleted. Though they were effective a few years ago as they employ technologies that were not originally meant for tracking purpose, later versions of those technologies have made them partially ineffective. Nevertheless, the partial successes can still be utilized to create the tracking solutions more robust. Table 1 shows the status of current relevance.

4.1. *Experiment using local storage*

HTTP-cookie usage was disabled in this experiment. Our aim was to achieve similar or more reliable tracking capability results, as defined by the test parameters, without the use of HTTP-cookies. As the data stored in the local storage is not automatically sent back to the server, we need some extra effort to make it a part of the client–server communication. All the communication between a webserver and the local storage happens using a JavaScript file that is attached to each tracked

Table 1. Currency of the new technologies as tracking methods.

	Cookies	Local storage	ETags
Test 1	Success	Success	Success
Test 2	Success	Success	Success
Test 3	Success	Success	Success
Test 4	Success	Success	Success
Test 5	Fail	Fail	Fail
Test 6	Fail	Fail	Fail
Test 7	Fail	Fail	Partial Success

webpage. We tested the communication using asynchronous communication (AJAX) when the web page contents need to be customized dynamically for each visitor. When the unique identifier stored in the "Local Storage" is used only for tracking process, it was saved into a hidden field within the webform and sent to the server on the next post-back action.

4.2. *Experiment using entity tags (ETags)*

Unlike "Local storage", ETags are inherently a mode of communication between server and client browsers, like HTTP-cookies. The cookie usage was disabled for this experiment to simulate the tracking mechanism, by only using ETags.

Tracking-pixels were assigned with the URL of the tracking service. Click-pixels, Conversion-pixels and other tracking-pixels caused the client browsers to send an HTTP-request to the tracking server. As the first step, the server examines the headers for an "If-None-Match header, which if present indicate the existence of a tracking ETag. A unique identifier for each user was set as ETag, similar a cookie-based tracking process. If the request header "If-None-Match" is not found, it indicates the start of a new tracking process, in which case the server adds two new headers to the HTTP-response: "Cache-control" header enabling caching on client and "ETag" header with the visitor's unique identifier as the value.

The same ETag must be repeatedly set on every response during all subsequent communication between the webserver and the client browser. Else, a response without an ETag and Cache-Control header or a directive will cause the browser to not use the previous browser cache, thereby losing the tracking capability of the ETag.

4.3. *Business insights gathering experiment*

The above simulation environment setup for an AM network, allowed us to observe the insights gathering process within an e-commerce environment. While functioning as an AM tracking services provider, information was limited to gathered click-data and matching conversion-data. The tracking process identifies the user only by a unique numeric identifier. The IP address is unique during a session, but not over longer durations, depending on the IP address leasing period of the DHCP server. But the geographical location of the user is revealed by the IP address, which can be matched with browser language to reveal the possible nationality or ethnicity of the visitor. The first visit, and frequency of subsequent visits as well as successful monetary outcomes, total purchase values, purchase per visit ratios could be calculated using tracked data attributed to the unique identifier. By placing a tracking-pixel in every page of the tracked site, we were able to monitor how long the visitor spent on each page which allowed us to create information such as the most popular pages, the most logical order of navigation, dead-locks that would usually cause the visitor to leave the site, etc. Products perused, what category of products attracted the most attention and the outcome are important business insights for a marketeer.

At this level of tracking, despite knowing the approximate location, language and buying habits, the visitor is only known by a number, without any PII.

Without any further changes to hardware or technology involved, we could extend the knowledge of the visitor-habits further and create even more marketable information by extending our view beyond the tracked site. When the visitor interacts with any other e-commerce or affiliate sites, that are subscribed to the same tracking service, the "referrer" header of the HTTP request revealed the current domain name, while the unique identifier remains the same across all the visited domains. Any products that were perused at multiple domains gives away the current urging purchase desire of the visitor. The knowledge of non-purchase at one site, can be sold to the next site as a premium lead such as the "remarketing" leads provided by many such services. Increasing the number of tracked domains by one tracking services provider increases the details of a tracked user, thereby also increasing the amount of marketable information. PII were still not available at this level of tracking.

We extended our simulation setup by adding a new domain that was fully accessible to the tracking domain and introduced a homepage that required a user account to access the site. That led to the first level of personal privacy intrusion, as that enabled the tracking service to combine the anonymous user-persona that was well-developed using the above- mentioned information, with a real person identifiable with an email address. Names, addresses, affiliations or any other information could be gathered in this process, depending on the motivation to lead a user to provide additional data in exchange of services provided.

Instead of a local account that uses a user name and password, by offering the log-in facility with Facebook, Google, Microsoft, Twitter and similar authentication providers, we can extract information associated with the user-profile to further enrich the tracked persona with information that appear in social media platforms. While providing wider visibility of the persona, this provided a much higher level of privacy intrusion. Converting the service provided by this new site to a social-media site would allow us to gather even more multi-faceted information such as political, social, environmental views and activities, family members and their activities, recent places of visit, including exact current location, which indicates the highest level of privacy intrusion.

5. Discussion

Unlike HTTP-cookies, the state management methods discussed here are not by design, technologies invented for tracking purposes. Methods that automatically transfer persisted identifiers back to the webserver with each HTTP-request, without having to implement specific code for such functionality, is a good candidate for tracking purpose. It reduces the number of points of failure. By design, both HTTP-cookie and ETags fulfil this condition. Webservers set cookies or the ETags, and on subsequent requests look for the cookies (by the name) or the ETags (by the value).

It is the responsibility of the browser to return the unique identifier to the server, with every request. In case of "Local Storage" it is not designed to send its values back to the server. It is meant to be used by the code running on client browser. Therefore, additional efforts are required to extract the information from the local storage and post it back to the server.

The "super cookie" concept and associated technologies were not designed to be used for the purpose of tracking, therefore future developments and new releases of those technologies can inadvertently make them unusable for tracking. As a technology that formed the super cookie concept "Adobe Flash Local shared objects" commonly known as "Flash cookies" have been intentionally upgraded by Adobe, to prevent them from being used as tracking technologies. Further, most browsers have by default, disabled access to flash content and require user's explicit permission. Reference 6 found that ETag retained their identifier values even when the cookies were blocked in a browser and when using "Private browsing mode". Results of the above experiments show that all the browsers now block ETags and Local storage, in both of the above scenarios. Therefore, keeping abreast with current developments of these technologies will enable researchers to adapt to these changes and modify the techniques to stay ahead of these changing technologies.

However, as seen in the results in Table 1, tracking capabilities using "Local Storage" perform equally well as HTTP-cookie-based traditional tracking technologies. Most common browsers have visual indicators on the browser window to show the use of HTTP-cookies within a site. For example, Chrome has a small cookie icon at the end of the URL address bar at the top of the windows. On clicking it, even the least-tech savvy users can delete or even block the cookies to that specific site, thereby failing the tracking process within that browser completely. In contrary, the use of local storage is not as visible to the user; therefore, to view the data in the local storage, requires user to dig deeper, such as use the "Developer Tools" that are accessible to users with more technical sophistication. Nevertheless, deleting HTTP-cookies now deletes local storage too, in newer versions of modern browsers.

ETags have an advantage over the other two methods, as ETag values are meant for the caching engines and therefore not easily visible to the general user. Also, the tools that are readily accessible on the user interface to remove or block cookies, do not delete the ETags, though they affect both the HTTP-cookies and local storage. But by removing browsing data including cache history, all identifiers can be removed.

Though we have displayed how these methods could be used without using cookies, for tracking purpose, we do not consider them as alternatives for HTTP-cookies. We recommend using cookies as the primary means for tracking, while using other methods in combination to make the process more robust.

The experiments above also verified the often-unintended information breaches. Following data security breaches and privacy threats were simulated during following technology usage scenarios, which are common in personal and business environments.

5.1. *Tracking data spillage*

With the above setup, we were able to simulate the complete tracking process to demonstrate different affiliate marketing models; cost-per-click (CPC), cost-per-mille (CPM) and cost-per-acquisition (CPA). The range of information exposed to the tracking service provider was observed. E-commerce practitioners who subscribe to the services of an AMN expect the AMN to monitor only transactions belonging to affiliate-generated web traffic. Instead, as the tracking pixel is placed on the payment confirmation page and a confirmation page is sent to every customer at the end of a payment, it triggered the conversion tracking process for every transaction. This includes information related to visitors who came through organic searches, paid advertising, search-engine advertising and every other traffic generation method. The tracking server can easily differentiate the AM generated traffic from non-AM traffic by the presence of an accompanying HTTP-cookie, which has been placed by the tracking server during click-tracking process. In an AM scenario, all web traffic that does not have a tracking cookie will be ignored by the tracking server and classified as non-AM generated traffic. However, enterprises are unaware that the tracking service has the capability of capturing all online purchases of the subscribed e-commerce practitioners.

This information leakage worsens with the popular practice of using services such as "Google Tag Manager", where e-commerce sites link their pixel-code via the Tag Manager URL instead of triggering directly on the tracking server. This exposes all online sales data to two different service providers, both of whom could use that information to generate additional value-added services, that are useful for the marketing efforts of competitors. For instance, remarketing sales leads that are offered at a higher price are based on the information on unsuccessful sales at competitors' e-commerce sites, since tracking service providers have visibility over customer interactions within all sites that have subscribed to their services. As Google services have wider visibility across most of the internet, using a single customer identifier across all sites, each customer's online interactions can be easily linked up to create a comprehensive behavioral profile. Some business managers who are uninformed about the information security breaches and the associated disadvantages choose to ignore security risk over the convenience of analytics (when their sales data are combined with the rest of business analytics data).

Tracking process for business analytics requires a tracking pixel to be embedded in every webpage that needs tracking. This triggers a tracking event with each step of the way during a browsing session, allowing an enterprise to gather a rich set of behavioral data of their customers. With a single User ID feature Google's Universal Analytics can track a user across multiple devices (e.g. phone, tablet, laptop, desktop, etc.) and across all participating sites in to one browsing profile, which makes the data very insightful to a practitioner.[7,38] Google's Universal Analytics guidelines make end-user privacy policy explicitly a practitioner's responsibility. Their terms and conditions state: "When you implement Universal Analytics, it is your

responsibility to ensure that your use is legally compliant, including with any local or regional requirements for specific notification to users".[39]

5.2. *CDN exposure*

We created a service endpoint on tracking server to serve a JavaScript library simulating the common use of JavaScript libraries from public CDNs. Web pages were created within the Dev domain that had links to those JavaScript libraries within their headers. Some pages were setup to use "Local Storage" as tracking technology in place of HTTP-cookies.[9,16]

CDNs are popular among web developers to reduce network latency. It is also common practice to link to most of the popular JavaScript libraries, CSS files and font files through CDNs. A compromised JavaScript file can provide control and access to sensitive data within a page, or in "Local Storage" and user inputs. Our tests were able to steal the visitor IDs from Local Storage, hidden fields on forms, change DOM elements, etc. Other static content providing CDNs can be used to stuff cookies, as discussed in cookie stuffing fraud in AM.[2]

5.3. *Click-bait*

Some YouTube videos, social media posts and links to blogs on diverse topics that appear as non-advertising material to site visitors can deliver harmful content without warning or consent.[40] Those links can direct visitors to servers that may place tracking cookies, display or simulate "clicks" on advertisements and undertake similar nefarious activities invisible to the visitor. To simulate this scenario, multiple endpoints were created on the tracking server that serves pure HTML content to the caller. Each of the endpoint URLs were posted to multiple test machines in Dev test domain simulating posts in social media networks. When a post was read or a comment was added using a client browser, the endpoint serving the HTML content was able to place a "connex.net.nz" cookie on the client machine. The main requirement for a successful tracking process is to place a cookie with a unique identifier during the first visit of a user. During all subsequent visits to any websites tracked by the said tracking server, the cookie will be sent back to the tracking server, by the client browser, thus identifying itself.

This collated information is sufficient for Cambridge Analytica style service providers to offer targeted campaigns in multitude of areas such as sales campaigns, political, social, environmental campaigns, etc to influence people's opinions.[29] With the combination of the two above-discussed datasets, despite having a quite clear picture of the persona, we still cannot personally identify the person in real-world.

5.4. *OAuth*

With this test, we created an application with "oAuth" or OpenID style authorization and access delegation service, as often found using Facebook, Google, Twitter,

etc. Though applications can request access to additional features connected to the user account, such as their complete name, access to contact list, etc., even with only the basic information as login name, it is possible to give a name and a face to the hitherto anonymous but comprehensive digital persona that we have created previously.

5.5. *Privacy woes vs. technological necessities*

As users, privacy groups and countries are getting more concerned about privacy and user rights, more regional and local regulations such as GDPR.[13] are being implemented that restrict online tracking activities. Tracking activities can be divided into three categories and each category of tracking has different levels of privacy implications therefore should be addressed differently:

(1) Purely technical: Tracking process used in Affiliate Marketing as discussed above falls into this category. E-marketing methods necessitate the ability to track a visitor from the source of the web traffic generation up to completion of transaction. No personally identifiable information (PII) is gathered in the process, it only uses a unique identifier assigned to each user. A "click-pixel" in advertisement-carrying pages and one "conversion-pixel" in payment-confirmation page are the only tracking requirement for this kind of tracking service. This mode of tracking does not create privacy concerns to the users, therefore new regulations and policies need to consider the importance current and future technological needs of this category of tracking and state management.
(2) Non-PII-based: The tracking process used by business analytic services fall into this category. The data gathering process goes well beyond the sheer technical necessity for e-commerce, as more comprehensive behavioral information is gathered for marketing purposes.[18,41] Though the identity of the user is not known to the tracking service, a comprehensive digital persona can be created using the gathered behavioral information across the Internet. Service providers can act upon that information by displaying targeted advertisements or specific political, religious and social content to influence them as in the case of Cambridge Analytica.[29] This can be harmful and detrimental to the unsuspecting user.

    This category can span from harmless and non-privacy intrusive services to information-scavenging nefarious operators. At the lower end of the scale are the e-marketing tracking service providers such as AMNs mentioned in the previous group, but who may have sought to venture a little deeper into information gathering process than required to operate as a purely tracking technology operator. At the opposite end of the scale are entities gathering business intelligence who are operating closer to the boarder of the next group described. Web scraping and web crawling activities form an important part of their activities. They usually offer free services and tools, so that they can place tracking cookies

into the browsers of unsuspecting visitors. They may get people to sign up for a free service by filling out forms requesting personally identifiable data such as names, contact e-mails, etc or, they may even ask site-visitors to sign in with social media credentials, which would allow them to link the anonymous digital personas they have created with a real name and a face. These service providers usually take a bottom-up approach into user profile creation, which means, they first gather many pieces of behavioral information of people without knowing the exact identity of the person. When they have gathered sufficient data to create a digital persona that is seemingly unique, it will be attempted connect the real-world identity to the anonymous persona. One of the major differences between this category and PII-based category is that the service providers at this level do not have a product or service that has a global reach.

(3) PII based: Providers of this category are set apart from the other two categories due to one or more products or services that they have with a global reach. The global reach is important, because with that millions of customers around the globe will have an account with the provider. They will gather PIIs of the users at the time of opening their accounts with the provider. This will allow the provider to gather behavioral data on that person, over time. This is a top-down approach for user profile creation, where first the person is positively identified, and then over time, behavioral data is accumulated. Microsoft, Apple, Google, Facebook, Twitter, LinkedIn, and other social media companies fall into this category.[33,42] Most people have an account with one or more services of these tech giants.

A few other common characteristics of these providers are that their product offerings are delivered across multiple hardware platforms such as wearables, mobile phones, tablets, laptops and desktops. That provides uninterrupted connectivity to the user, and continuous tracking capability to the service provider. Microsoft, Apple and Google have the advantage of operating system level identity knowledge.[43] The next best preferable method would be browser level identification, through which a user's browsing data can be collected. Though popular browsers like Chrome can be used without logging into it, users may be aware, browser's continuous reminder to log-in to the browser, which makes tracking easier for the browser manufacturer. Even if tech-savvy users know that they are constantly tracked while being logged-in to these tech giants, users choose to stay logged in, due to convenience.

6. Future Direction

It can be rightfully expected that any future developments to state-management technologies such as HTTP cookies would still adhere to the requirement of XDT. Further research efforts could increase the robustness of the tracking technology by supplementing existing cookie-based tracking technology with alternative non-traditional technologies. In this research, we have used stateful tracking technologies. It will be useful to investigate how stateless tracking technologies can add to the

robustness of the above tracking methods.[8,9,16] As those alternative technologies keep changing their capability to be used as a tracking technology, continuous research efforts are needed to adapt to those changes that can drive the efficacy of cross-domain tracking capabilities.

Policy developers need to consider each of these tracking scenario categories individually and holistically during policy development, as more and more countries are currently developing country and region-specific regulations. European GDPR has frustrated and caused click-fatigue among users, that many have been often clicking "accept" for want of a better option.[14] If the policies are not well thought out, it only adds to more bureaucracy without achieving intended results.[12]

Further research could be carried out to develop a framework that will translate higher level privacy requirements in layman's terms to pre-agreed technical implementation categories, that should be implemented during the negotiation of the connection between the web server and the client browser. Though every browser has a settings page, the current settings do not translate to universally accepted technical definitions. User's choice of tracking capabilities based on the capabilities and categories discussed above can be translated to specific implementations. When adhered to by web servers, browsers and web application developers, click-fatigue can be avoided, which is due to the current necessity to make those choices on a site-by-site basis, at every site.

References

1. D. M. Kristol and L. Montulli, HTTP state management mechanism, *IETF Internet RFCs* **2109** (1997), https://tools.ietf.org/html/rfc2109.
2. B. R. Amarasekara and A. Mathrani, Exploring risk and fraud scenarios in affiliate marketing technologies from the advertiser's perspective, in *Proc. Australasian Conf. Information Systems (ACIS2015)* (Adelaide, 2015).
3. N. Chachra, Understanding URL abuse for profit, *Doctoral Dissertation* (University of California, San Diego, CA, 2015).
4. B. Edelman and W. Brandi, Risk, information, and incentives in online affiliate marketing, *J. Market. Res.* **LII** (2015) 1–12.
5. P. Snyder and C. Kanich, No Please, After You: Detecting fraud in affiliate marketing networks, in *Proc. Workshop Economics of Information Security (WEIS)* (University of Illinois, 2015).
6. M. D. Ayenson, D. J. Wambach, A. Soltani, N. Good and C. J. Hoofnagle, *Flash Cookies and Privacy II: Now with HTML5 and ETag Respawning* 2011. DOI: https://dx.doi.org/10.2139/ssrn.1898390.
7. R. Binns, U. Lyngs, M. Van Kleek, J. Zhao, T. Libert and N. Shadbolt, Third party tracking in the mobile ecosystem, in *Proc. WebSci'18* (ACM: Amsterdam, Netherlands, 2018), pp. 23–31.
8. P. Eckersley, How unique is your web browser?, in *Proc. Privacy Enhancing Technologies* (Springer, 2010)
9. P. Laperdrix, W. Rudametkin and B. Baudry, Beauty and the Beast: Diverting modern web browsers to build unique browser fingerprints, in *Proc. 37th IEEE Sump. Security and Privacy* (San Jose, 2016).

10. A. Soltani, S. Canty, Q. Mayo, L. Thomas and C. J. Hoofnagle, Flash cookies and privacy, in *Proc. AAAI Spring Symp. Intelligent Information Privacy Management* (Palo Alto, California, 2010), pp. 158–163.
11. W3C, W3C Recommendtion-Web Storage, Available at https://www.w3.org/TR/2013/REC-webstorage-20130730/(2013).
12. S. Wachter and B. Mittelstadt, A right to reasonable inferences: Re-thinking data protection law in the age of big data and AI, *Colomb. Bus. Law Rev.* **2019**(2) (2019) 130.
13. GDPR, General data protection regulation, in *Official Journal of the European Union* (2016).
14. C. Utz, M. Degeling, S. Fahl, F. Schaub and T. Holz, (Un)informed consent: Studying GDPR consent notices in the field, in *Proc. ACM SIGSAC Conf. Computer and Communications Security (CCS'19)* (London, UK, ACM, New York, 2019), p. 18.
15. B. R. Amarasekara, A. Mathrani and C. Scogings, Improving the robustness of the cross-domain tracking process (Springer, Singapore, 2020), pp. 260–270.
16. S. Englehardt and A. Narayanan, Online tracking: A 1-million-site measurement and analysis, in *Proc. Proceedings of the ACM SIGSAC Conference on Computer and Communications Security* (Association for Computing Machinery, Vienna, Austria, 2016).
17. R. Fielding and J. Reschke, Hypertext Transfer Protocol (HTTP/1.1): Conditional Requests, *IETF Internet RFCs* **7232** (2014), https://tools.ietf.org/html/rfc7232.
18. A. Baumann, J. Haupt, F. Gebert and S. Lessmann, The price of privacy: An evaluation of the economic value of collecting clickstream data, *Business & Information Systems Engineerng* **61**(4) (2019) 413–431.
19. N. Chachra, S. Savage and G. M. Voelker, Affiliate crookies: Characterizing affiliate marketing abuse, in *Proc. IMC '152015 ACM Conf. Internet Measurement Conf.* (ACM, Tokyo, Japan, 2015), pp. 41–47.
20. R. Olbrich, P. M. Bormann and M. Hundt, Analyzing the click path of affiliate-marketing campaigns: Interacting effects of affiliates' design parameters with merchants' search-engine advertising, *J. Advert. Res.* **59**(3) (2019) 342–356.
21. P. Snyder and C. Kanich, Characterizing fraud and its ramifications in affiliate marketing networks, *J. Cybersecur.* **2**(1) (2016) 71–81.
22. B. R. Amarasekara and A. Mathrani, Revenue fraud in e-commerce platforms: Challenges and solutions for affiliate marketing, in *Cyber Security and Policy: A Substantive Dialogue*, eds. A. Colarik, J. Jang-Jaccard and A. Mathrani (Massey University Press: Auckland, New Zealand, 2017), pp. 67–87.
23. D. Brear and S. J. Barnes, Assessing the value of online affiilate marketing in the UK financial services industy, *Int. J. Electron. Fin.* (2008).
24. A. Norouzi, An integrated survey in affiliate marketing network, in *Proc. 2nd World Conference on Technology, Innovation and Entrepreneurship* (Istanbul, Turkey, 2017), pp. 299–309.
25. K. Pawan and S. Gursimranjit, Using social media and digital marketing tools and techniques for developing brand equity with connected consumers, in *Handbook of Research on Innovations in Technology and Marketing for the Connected Consumer*, ed. D. Sumesh Singh (IGI Global, Hershey, PA, USA, 2020), pp. 336–355.
26. S. A. Suryanarayana, D. Sarne and S. Kraus, Information disclosure and partner management in affiliate marketing, in *Proc. First Int. Conf. Distributed Artificial Intelligence* (ACM, Beijing, China, 2019), pp. 1–8.
27. T. Libert, Exposing the invisible web: An analysis of third-party HTTP requests on 1 million websites, *International Journal of Communication* (2015).

28. C. J. Hoofnagle, J. Urban and S. Li, Privacy and modern advertising: Most US internet users want "Do Not Track" to stop collection of data about their online activities, in *Proc. Amsterdam Privacy Conf.* (Amsterdam, Netherlands, 2012).
29. A. Richterich, How data-driven research fuelled the cambridge analytica controversy, *Open J. Sociopolit. Stud.* **11**(2) (2018) 528–543.
30. P. O'Brien, S. W. H. Young, K. Arlitsch and K. Benedict, Protecting privacy on the web: A study of HTTPS and google analytics implementation in academic library websites, *Online Inf. Rev.* **42**(6) (2018) 734–751.
31. S. Schelter and J. Kunegis, Tracking the trackers: A large-scale analysis of embedded web trackers, in *Proc. AAAI Int. Conf. Weblogs and Social Media* (Cologne, Germany, 2016).
32. A. Narayanan and D. Reisman, The Princeton web transparency and accountability project, in *Transparent Data Mining for Big and Small Data* (Springer, 2017), pp. 45–67.
33. B. Krishnamurthy and C. E. Wills, Privacy diffusion on the web: A longitudinal perspective, in *Proc. WWW'09-18th Int. World Wide Web Conf.* (Madrid, Spain, 2009), pp. 541–550.
34. B. R. Amarasekara, Analysis, design and simulation of fraud and vulnerability management in affiliate marketing, Master Thesis, Massey University of Auckland (2017).
35. A. R. Hevner, S. T. March, J. Park and S. Ram, Design science in information systems research, *MIS Quart.* **28**(1) (2004) 75–105.
36. S. T. March and G. Smith, Design and natural science research on information technology, *Decis. Support Syst.* **15**(4) (1995) 251–266.
37. J. Nunamaker, M. Chen and T. D. M. Pruding, Systems development in information systems research, *J. Manage. Inf. Syst.* **7**(3) (1991) 89–106.
38. Google, User-ID limits, cited 30 September 2019, Available at https://support.google.com/analytics/answer/3123668?hl=en.
39. Google, Universal Analytics usage guidelines, cited 30 September 2019, Available at https://support.google.com/analytics/answer/2795983?hl=en.
40. A. Mathur, A. Narayanan and M. Chetty, Endorsements on social media: An empirical study of affiliate marketing disclosures on youtube and pinterest, in *Proc. ACM Human–Computer Interaction* (2018).
41. A. Lerner, A. K. Simpson, T. Kohno and F. Roesner, Internet jones and the raiders of the lost tracker: An archaelogical study of web tracking from 1996 to 2016, in *Proc. 25th USENIX Security Symp. (USENIX Security 16)* (Austin, TX, 2016).
42. J. R. Mayer and J. C. Mitchell, Third-party web tracking: Policy and technology, in *Proc. IEEE Symp. Security and Privacy* (2012), pp. 413–427.
43. J. Gamba, M. Rashed, A. Razaghpanah, J. Tapiador and N. Vallina-Rodriguez, *An Analysis of Pre-installed Android Software* (2019).

Online Tracking: When Does it Become Stalking?

Amarasekara, B

2021-05-25